

Supplementary Information

Supplementary Table 1. List of genes commonly expressed in all single cells.

Symbol	Name
Actb	Actin, beta, cytoplasmic
Actg1	Actin, gamma, cytoplasmic 1
Atp5e	ATP synthase, H ⁺ transporting, mitochondrial F1 complex, epsilon subunit
Bsg	Basigin
Ctsh	Cathepsin H
Ddx5	DEAD (Asp-Glu-Ala-Asp) box polypeptide 5
Dstn	Destrin
Eef1a1	Eukaryotic translation elongation factor 1 alpha 1
Eif1	Eukaryotic translation initiation factor 1
Fau	Finkel-Biskis-Reilly murine sarcoma virus (FBR-MuSV) ubiquitously expressed (fox derived)
Fstl1	Follistatin-like 1
Fth1	Ferritin heavy chain 1
Ftl1	Ferritin light chain 1
Gm6251	Predicted gene 6251
Gm6654	Predicted gene 6654
Gm9320	Predicted gene 9320
Gng5	Guanine nucleotide binding protein (G protein), gamma 5 subunit
H3f3b	H3 histone, family 3B
Lars2	Leucyl-tRNA synthetase, mitochondrial
Malat1	Metastasis associated lung adenocarcinoma transcript 1 (non-coding RNA)
Myl6	Myosin, light polypeptide 6, alkali, smooth muscle and non-muscle
Ndufa3	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 3
Ppia	Peptidylprolyl isomerase A
Rn45s	45S pre-ribosomal RNA
Rpl10	Ribosomal protein 10
Rpl10a	Ribosomal protein L10A
Rpl11	Ribosomal protein L11
Rpl13	Ribosomal protein L13
Rpl13a	Ribosomal protein L13a
Rpl17	Ribosomal protein L17
Rpl18a	Ribosomal protein L18A
Rpl21	Ribosomal protein L21
Rpl23	Ribosomal protein L23
Rpl23a	Ribosomal protein L23a
Rpl24	Ribosomal protein L24
Rpl27a	Ribosomal protein L27a
Rpl31-ps12	Ribosomal protein L31, pseudogene 12
Rpl32	Ribosomal protein L32
Rpl35	Ribosomal protein L35
Rpl35a	Ribosomal protein L35a
Rpl36	Ribosomal protein L36
Rpl37	Ribosomal protein L37
Rpl37a	Ribosomal protein L37a
Rpl39	Ribosomal protein L39
Rpl4	Ribosomal protein L4
Rpl41	Ribosomal protein L41
Rpl7	Ribosomal protein L7
Rpl8	Ribosomal protein L8
Rpl9	Ribosomal protein L9
Rplp0	Ribosomal protein, large, P0
Rplp1	Ribosomal protein, large, P1
Rplp2	Ribosomal protein, large P2
Rplp2-ps1	Ribosomal protein, large P2, pseudogene 1
Rps14	Ribosomal protein S14
Rps15	Ribosomal protein S15
Rps17	Ribosomal protein S17
Rps18	Ribosomal protein S18
Rps19	Ribosomal protein S19
Rps19-ps3	Ribosomal protein S19, pseudogene 3
Rps23	Ribosomal protein S23
Rps24	Ribosomal protein S24
Rps25	Ribosomal protein S25
Rps26	Ribosomal protein S26
Rps27a	Ribosomal protein S27a
Rps28	Ribosomal protein S28
Rps3a	Ribosomal protein S3a
Rps4x	Ribosomal protein S4, X-linked
Rps5	Ribosomal protein S5

Continuation of Supplementary Table 1. List of genes commonly expressed in all single cells.

Symbol	Name
Rps6	Ribosomal protein S6
Rps7	Ribosomal protein S7
Rps8	Ribosomal protein S8
Rps9	Ribosomal protein S9
S100a11	S100 calcium binding protein A11 (calgizzarin)
Sod1	Superoxide dismutase 1, soluble
Tmsb4x	Thymosin, beta 4, X chromosome
Tpt1	Tumor protein, translationally-controlled 1
Uba52	Ubiquitin A-52 residue ribosomal protein fusion product 1
Ubb	Ubiquitin B
Wbp5	WW domain binding protein 5
Ywhae	Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, epsilon polypeptide
Ywhaz	Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, zeta polypeptide

Supplementary Table 2. List of DELTAgene™ Assays (Fluidigm) used for microfluidic single cell multiplexed qPCR experiments on embryonic mouse lung epithelial cells.

Gene Symbol	Gene Full Name	DELTAgene Assay ID	comment
Positive Control			
Actb	actin, beta	GEP00055164	housekeeping
Gapdh	glyceraldehyde-3-phosphate dehydrogenase	GEP00055172	housekeeping
Cdh1	cadherin 1	GEP00055166	epithelial
Proliferation marker			
Mki67	antigen identified by monoclonal antibody Ki 67	GEP00055177	proliferation
Birc5	baculoviral IAP repeat-containing 5	GEA00012818	proliferation
Cdk6	cyclin-dependent kinase 6	GEA00012826	proliferation
Jub	ajuba	GEA00017912	proliferation
Top2a	topoisomerase (DNA) II alpha	GEA00012950	proliferation
Negative control			
Ascl1	achaete-scute complex homolog 1 (Drosophila)	GEA00012171	neuroendocrine
Dll1	delta-like 1 (Drosophila)	GEA00008057	neuroendocrine
Ptpnc	protein tyrosine phosphatase, receptor type, C	GEA00013196	leukocytes
Trp63	transformation related protein 63	GEP00055181	basal
Marker genes for alveolar and bronchiolar cell types			
Abca3	ATP-binding cassette, sub-family A (ABC1), member 3	GEA00050074	AT2
Sftpb	surfactant associated protein B	GEA00031945	AT2
Muc1	mucin 1, transmembrane	GEA00012891	AT2
Sftpc	surfactant associated protein C	GEA00031944	AT2
Lyz2	lysozyme 2	GEP00055184	AT2
Aqp5	aquaporin 5	GEA00037602	AT1
Pdpn	podoplanin	GEA00040971	AT1
Ager	advanced glycosylation end product-specific receptor	GEA00045168	AT1
Foxj1	forkhead box J1	GEA00036645	Ciliated
Scgb1a1	secretoglobulin, family 1A, member 1 (uteroglobulin)	GEA00037187	Clara
Other genes (enriched in AT1/AT2, stemness, signaling, chromatin, extracellular matrix, disease, cancer)			
Ap3b1	adaptor-related protein complex 3, beta 1 subunit	GEA00013340	disease
Atf3	activating transcription factor 3	GEA00017978	AT2
Atf6	activating transcription factor 6	GEA00047614	disease
Axin2	axin2	GEA00023066	signaling
Bmi1	Bmi1 polycomb ring finger oncogene	GEP00055165	stemness
Bmp3	bone morphogenetic protein 3	GEA00012010	signaling
Bmpr1a	bone morphogenetic protein receptor, type 1A	GEA00013341	signaling
Bptf	bromodomain PHD finger transcription factor	GEA00018244	chromatin
Brd4	bromodomain containing 4	GEA00018438	chromatin
Cav1	caveolin 1, caveolae protein	GEA00012822	signaling
Cdkn1a	cyclin-dependent kinase inhibitor 1A (P21)	GEA00011737	signaling
Cebpa	CCAAT/enhancer binding protein (C/EBP), alpha	GEP00055168	AT2
Cfr	cystic fibrosis transmembrane conductance regulator	GEA00043570	disease
Cic	capicua homolog (Drosophila)	GEA00018361	cancer
Cldn3	claudin 3	GEA00007979	AT2
Col3a1	collagen, type III, alpha 1	GEA00050479	ECM
Dnmt3a	DNA methyltransferase 3A	GEA00012143	chromatin
Ect2	ect2 oncogene	GEP00055183	cancer
Efnal	ephrin A1	GEA00039228	cancer
Egfr	epidermal growth factor receptor	GEA00009844	cancer
Egr1	early growth response 1	GEA00011750	AT2
ErbB2	v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived	GEP00055169	signaling
Ezh2	enhancer of zeste homolog 2 (Drosophila)	GEA00013271	chromatin
Foxa2	forkhead box A2	GEP00055170	AT2
Foxj1	forkhead box P1	GEP00055171	AT2
Foxp2	forkhead box P2	GEA00013470	AT2
Gata6	GATA binding protein 6	GEA00011865	development

Continuation of Supplementary Table 2. List of DELTAgene™ Assays (Fluidigm) used for microfluidic single cell multiplexed qPCR experiments on embryonic mouse lung epithelial cells.

Gene Symbol	Gene Full Name	DELTAgene Assay ID	comment
Hes1	hairy and enhancer of split 1 (Drosophila)	GEA00011756	signaling
Hgf	hepatocyte growth factor	GEA00048519	ECM
Hhex	hematopoietically expressed homeobox	GEA00018343	development
Hps1	Hermansky-Pudlak syndrome 1 homolog (human)	GEA00048735	disease
Id1	inhibitor of DNA binding 1	GEA00011760	AT2
Id2	inhibitor of DNA binding 2	GEP00055173	signaling
Inhba	inhibin beta-A	GEA00012170	development
Itga6	integrin alpha 6	GEA00013287	stemness
Itgb4	integrin beta 4	GEP00055185	stemness
Jag1	jagged 1	GEA00012064	signaling
Jun	Jun oncogene	GEP00055174	cancer
Junb	Jun-B oncogene	GEA00018493	AT2
Keap1	kelch-like ECH-associated protein 1	GEA00040802	cancer
Kit	kit oncogene	GEA00011823	stemness
Klf4	Kruppel-like factor 4 (gut)	GEA00013370	stemness
Klf9	Kruppel-like factor 9	GEA00012151	AT2
Kras	v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog	GEA00012872	cancer
Krt5	keratin 5	GEP00055175	stemness
Lama5	laminin, alpha 5	GEA00050572	ECM
Lefty1	left right determination factor 1	GEA00011859	signaling
Lgr4	leucine-rich repeat-containing G protein-coupled receptor 4	GEA00008764	stemness
Lgr6	leucine-rich repeat-containing G protein-coupled receptor 6	GEA00008838	stemness
Lrig1	leucine-rich repeats and immunoglobulin-like domains 1	GEA00048532	stemness
Mbd3	methyl-CpG binding domain protein 3	GEA00022636	chromatin
Mll1	myeloid/lymphoid or mixed-lineage leukemia 1	GEA00012881	chromatin
Mmp11	matrix metalloproteinase 11	GEA00043111	AT1
Myc	myelocytomatosis oncogene	GEA00011771	cancer
Mycl1	v-myc myelocytomatosis viral oncogene homolog 1, lung carcinoma derived (avian)	GEA00012894	cancer
Nanog	Nanog homeobox	GEA00051980	stemness
Notch1	notch 1	GEA00011772	signaling
Nrg1	neuregulin 1	GEP00055178	signaling
Pax8	paired box gene 8	GEA00046089	development
Pdx1	pancreatic and duodenal homeobox 1	GEA00011847	development
Prdm16	PR domain containing 16	GEA00018416	signaling
Raf1	v-raf-leukemia viral oncogene 1	GEA00011891	signaling
Sftpd	surfactant associated protein D	GEA00013315	AT2
Shh	sonic hedgehog	GEA00012111	signaling
Smarca4	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily	GEA00022815	chromatin
Sox2	SRY-box containing gene 2	GEP00055179	stemness
Spry2	sprouty homolog 2 (Drosophila)	GEA00035957	signaling
Stat3	signal transducer and activator of transcription 3	GEA00003674	signaling
Syt7	synaptotagmin VII	GEA00044319	AT2
Tert	telomerase reverse transcriptase	GEA00011787	cancer
Tnc	tenascin C	GEP00055180	ECM
Trp53	transformation related protein 53	GEA00011788	cancer
Nkx2-1	Thyroid transcription factor 1/ NK2 homeobox 1	GEA00017957	AT2
Vegfa	vascular endothelial growth factor A	GEA00012621	AT1

Supplementary Data

Supplementary Data 1: Alignment statistics for all single cells with sequenced transcriptome.

Table listing all single cells, whose transcriptome was sequenced with a unique cell identifier as well as with number of raw reads and percentage of mapped reads obtained using the SAMtools flags¹.

Supplementary Data 2: R script used to analyze single cell RNA-seq data.

Supplementary Data 3: Single cell RNA-seq expression data for all 80 lung epithelial cells at E18.5.

Log₂(FPKM) values are given for all genes and for all 80 single lung epithelial cells isolated at E18.5 and discussed in Figure 1 and 2 as well as the positive bulk control samples for the same time point. In addition to the unique cell name (column 1), the time point of cell isolation (column 2) and the sample type (column 3), the putative cell type of each cell is shown as identified by PCA and hierarchical clustering analysis (see Figure 1b-c, Figure 2a and Methods).

Supplementary Data 4: Characterization of novel marker genes within defined epithelial cell populations of the lung.

Sheet 1: Shown are the top 30 genes for each cell population identified by the strongest correlation with the theoretical "perfect marker gene" for the respective cell population (Methods) as well as by

the lowest p-values calculated using Guilt-By-Association (Methods). These populations were identified as AT2, AT1, Clara, or Ciliated cells based on canonical marker genes (blue) and known marker genes (lighter blue). Genes were ranked as follows: (1) gene expression reported in the lung with limited or no spatial or cell-type information, (2) gene expression reported in the lung with spatial information or cell-type information, (3) lung-specific functional information has been reported about this gene *in vitro*, (4) lung-specific functional information has been reported about this gene *in vivo*. Representative publications that report lung-specific expression and mutant phenotypes are identified by PubMed Unique Identifier. Expression observations from Eurexpress Consortium data (<http://www.eurexpress.org/>) are identified by euxassay id² and results from the BioGPS project (<http://biogps.org>) are identified by Entrez Gene ID³. A brief description of these results has been included.

Sheet 2: All genes with a correlation coefficient greater than 0.3 concerning correlation with the theoretical "perfect marker gene" for the respective cell population are presented together with p-values calculated by Guilt-By-Association (Methods).

Supplementary Data 5: Gene ontology and KEGG pathway enrichment analysis (DAVID informatics Resources 6.7^{4,5}) of cell type specific genes for AT1, AT2, Clara and Ciliated cells as identified in the single cell RNA-seq data at E18.5.

Supplementary Data 6: Table listing genes identified by PCA to describe the variation in the data set of all *Sftpc*⁺ cells across 4 different time points (Figure 4).

Sheet 1, gene PC loadings: The gene names are shown together with Pearson correlation coefficient and p-value describing the correlation of each gene with one of the first three principal components. Genes with highest PC loading were used for hierarchical clustering in Figure 4a are shown in yellow.

Sheet 2, gene groups: Clustering of genes into 5 different groups based on hierarchical clustering shown in Figure 4a.

Sheet 3-7, GO group I-V: Gene ontology enrichment analysis for each individual gene group.

1. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
2. Diez-Roux, G. *et al.* A high-resolution anatomical atlas of the transcriptome in the mouse embryo. *Plos Biol* **9**, e1000582 (2011).
3. Wu, C. *et al.* BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol* **10**, R130 (2009).
4. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research* **37**, 1–13 (2009).
5. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4**, 44–57 (2009).